Rademacher Complexity in Deep Networks

Presenter: Haoyu Wei

May 2021

Recall

For given loss function ℓ,

- population risk: $\mathcal{R}(f) = \mathrm{E}l(f(x), y)$.
- empirical risk: $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$.

Given a training algorithm's choice $\widehat{f}$ in class $\mathcal{F}$, as well as some reference solution $\bar{f} \in \mathcal{F}$, we have the following decomposition

$$\mathcal{R}(\widehat{f}) = \underbrace{\mathcal{R}(\widehat{f}) - \widehat{\mathcal{R}}(\widehat{f})}_{\text{generalization}} + \underbrace{\widehat{\mathcal{R}}(\widehat{f}) - \widehat{\mathcal{R}}(\bar{f})}_{\text{optimization}} + \underbrace{\widehat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f})}_{\text{concentration}} + \underbrace{\mathcal{R}(\bar{f})}_{\text{approximation}} .$$

So for generalization, we want to control the following **uniform deviations**:

$$\sup_{f \in \mathcal{F}} \left(\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\right),$$

since $\widehat{f}$ is random.

Intuition

Example 18.1 ***(finite classes)*** As an example of what is possible, suppose we have $\mathcal{F} = (f_1, \ldots, f_k)$, meaning a finite function class $\mathcal{F}$ with $|\mathcal{F}| = k$. If we apply Hoeffding's inequality to each element of $\mathcal{F}$ and then union bound, we get, with probability at least $1 - \delta$, for every $f \in \mathcal{F}$,

$$\Pr[f(X) \neq Y] - \widehat{\Pr}[f(X) \neq Y] \leq \sqrt{\frac{\ln(k/\delta)}{2n}} \leq \sqrt{\frac{\ln|\mathcal{F}|}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Rademacher complexity will give us a way to replace $\ln|\mathcal{F}|$ in the preceding finite class example with something non-trivial in the case $|\mathcal{F}| = \infty$.

Definition 18.1 ***(Rademacher complexity)*** Given a set of vectors $V \subseteq \mathbb{R}^n$, define the **(unnormalized) Rademacher complexity** as

$$\mathrm{URad}(V) := \mathbb{E} \sup_{u \in V} \langle \epsilon, u \rangle, \qquad \mathrm{Rad}(V) := \frac{1}{n}\mathrm{URad}(V),$$

where $\mathbb{E}$ is uniform over the corners of the hyerpcube over $\epsilon \in \{\pm 1\}^n$ (each coordinate ϵ_i is a *Rademacher random variable*, meaning $\Pr[\epsilon_i = +1] = \frac{1}{2} = \Pr[\epsilon_i = -1]$, and all coordinates are iid).

Intuition

A close look at Example 18.1 in the lecture note.

Example (18.1)

Suppose $\mathcal{F} = \{f_1, \ldots, f_k\}$ is a finite function class with $|\mathcal{F}| = k$, denote $Z_i(f) := \mathbb{I}(f(X_i) \neq Y_i) \in \{0, 1\}$. Consider the 0-1 loss $\ell(f) = \mathbb{I}(f(X) \neq Y)$,

$$\sup_{f\in\mathcal{F}} \left(\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f)\right) = \sup_{f\in\mathcal{F}} \left[\mathrm{E}Z_i(f) - \frac{1}{n}\sum_{i=1}^{n} Z_i(f)\right] \leq \sum_{j\in[k]} \left[\mathrm{E}Z_i(f_j) - \frac{1}{n}\sum_{i=1}^{n} Z_i(f_j)\right].$$

On the other hand, by Hoeffding inequality for every $\epsilon > 0$,

$$\mathrm{P}\left[\sup_{f\in\mathcal{F}} \left(\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f)\right) \geq k\epsilon\right] \leq \sum_{j=1}^{k} \mathrm{P}\left(\mathrm{E}Z_{ji} - \frac{1}{n}\sum_{i=1}^{n} Z_{ji} \geq \epsilon\right) \leq k\exp\left(-\frac{2n^2\epsilon^2}{2n}\right)$$

take $\epsilon = \sqrt{\frac{\log(k/\delta)}{2n}}$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left(\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f)\right) \leq \sqrt{\frac{\log(k/\delta)}{2n}} \leq \sqrt{\frac{\log|\mathcal{F}|}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Intuition

When $|\mathcal{F}| = \infty$, we need an alternative!

Consider any loss function ℓ, denote $S := \{(X_i^\top, Y_i)^\top\}_{i=1}^n$ and the IID copy $S' := \{(X_i'^\top, Y_i')^\top\}_{i=1}^n$. By concentration inequality, we can first control the expectation of $\sup_{f\in\mathcal{F}}(\mathcal{R}(f) - \widehat{\mathcal{R}}(f))$.

$$
\begin{aligned}
\mathrm{E}\sup_{f\in\mathcal{F}}\left(\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f)\right) &= \mathrm{E}_S \sup_{f\in\mathcal{F}}\left(\mathrm{E}_{S'}\widehat{\mathcal{R}}_{S'}(f) - \widehat{\mathcal{R}}_S(f)\right) \\
&\le \frac{1}{n}\mathrm{E}_{S,S'}\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^n \left(\ell(f(X_i'), Y_i') - \ell(f(X_i), Y_i)\right)\right]^1 \\
&= \frac{1}{n}\mathrm{E}_{S,S'}\mathrm{E}_\epsilon \sup_{f,\, f'\in\mathcal{F}}\sum_{i=1}^n \epsilon_i\left(\ell(f'(X_i'), Y_i') - \ell(f(X_i), Y_i)\right) \\
&\le \frac{2}{n}\mathrm{E}_S\mathrm{E}_\epsilon \sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i \ell(f(X_i), Y_i)) = \frac{2}{n}\mathrm{E}_S\mathrm{E}_\epsilon \sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i\cdot \ell\circ f(Z_i),
\end{aligned}
$$

where $Z = (X^\top, Y)^\top$.

[1] some thing like the gap between test error and training error.

Intuition

Now we have obtained the inequalities between

$$\mathrm{E}\sup_{f\in\mathcal{F}}\Big(\mathcal{R}_\ell(f)-\widehat{\mathcal{R}}_\ell(f)\Big) \qquad \text{and} \qquad \mathrm{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot\big(\ell\circ f(Z_i)\big),$$

and hence

$$\sup_{f\in\mathcal{F}}\Big(\mathcal{R}_\ell(f)-\widehat{\mathcal{R}}_\ell(f)\Big) \qquad \text{and} \qquad \mathrm{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot\big(\ell\circ f(Z_i)\big),$$

Based on this intuition, we can purpose the Rademacher Complexity.

Rademacher Complexity[2]

Definition

Given a set of vectors $V \subseteq \mathbb{R}^n$, define the (un-normalized) *Rademacher complexity* as

$$\mathrm{URad}(V) := \mathrm{E} \sup_{u \in V} \langle \epsilon, u \rangle, \qquad \mathrm{Rad}(V) = \frac{1}{n} \mathrm{URad}(V),$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ with iid component $\mathrm{P}(\epsilon_i = +1) = \mathrm{P}(\epsilon_i = -1) = 1/2$.

Consider the sample $S = \{z_i\}_{i=1}^n$, the function class then is

$$\mathcal{F}_{|S} := \left\{ (f(z_1), \ldots, f(z_n))^\top : f \in \mathcal{F} \right\} \subseteq \mathbb{R}^n.$$

Within this definition,

$$\mathrm{URad}(\mathcal{F}_{|S}) = \mathrm{E}_\epsilon \sup_{u \in \mathcal{F}_{|S}} \langle \epsilon, u \rangle = \mathrm{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(z_i).$$

Specially, what we are most interested is

$$\mathrm{URad}\left((\ell \circ \mathcal{F})_{|S} \right) = \mathrm{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i \cdot (\ell \circ f)(Z_i).$$

[2] 24

Rademacher Complexity for Finite Class Bounds

Theorem (Massart Finite Lemma)

$$\mathrm{URad}(V) \leq \sup_{u \in V} \|u\|_2 \sqrt{2 \log |V|}.$$

This theorem exactly explain to us that Rademacher complexity $\mathrm{URad}(V)$ is a reasonable alternative of $\log |V|$ even in finite case. And this can be proved by the the following properties of sub-Gaussian variables.

Lemma (sub-Gaussian Properties)

If $X_i \sim \mathrm{subG}(c_i^2)$, *then* $\sum_{i=1}^n X_i \sim \mathrm{subG}(\|c\|_2^2)$, *and* $\mathrm{E} \max_{i \in [n]} X_i \leq \|c\|_2 \sqrt{2 \log n}$

The first result of this lemma is a direct result of the definition of sub-Gaussian variables, and the second result comes from the inequality

$$\mathrm{E} \max_{i \in [n]} X_i = \inf_{t > 0} \frac{1}{t} \mathrm{E} \log \max_{i \in [n]} \exp(t X_i) \leq \frac{1}{t} \mathrm{E} \log \sum_{i=1}^n \exp(t X_i) \leq \frac{1}{t} \mathrm{E} \log \sum_{i=1}^n \exp(t^2 c_i^2 / 2),$$

with its minimizer $t = \sqrt{2 \log n} / \|c\|_\infty$.

The Proof of Massart Finite Lemma

For any fixed $u \in V$, define $X_{u,i} := \epsilon_i u_i$ and $X_u = \sum_{i=1}^{n} X_{u,i} = \langle \epsilon, u \rangle$.

By Hoeffding Lemma[3], $X_{u,i} \sim \mathrm{subG}(u_i^2)$, hence $X_u \sim \mathrm{subG}(\|u\|_2^2)$, and by the previous Lemma and note that $|V|$ is finite, we have

$$\mathrm{URad}(V) = \mathrm{E}_\epsilon \max_{u \in V} X_u \leq \max_{u \in V} \|u\|_2 \sqrt{2 \log |V|}.$$

This is Massart Finite Lemma.

[3] If a centralized random variable $\xi \in [a, b]$, then $\xi \sim \mathrm{subG}((b-a)^2/4)$. The details can be seen in P36.

Some Remarks

Remark

$\mathrm{URad}(V)$ *to measure how "big" or "complicated" V is, we give the following sanity checks:*

1. $\mathrm{URad}(\{u\}) = \mathrm{E}\langle \epsilon, u \rangle = 0$, *as* $|V| = 1$ *is simple.*
2. *If* $V \subseteq V'$, *then* $\mathrm{URad}(V) \leq \mathrm{URad}(V')$.
3. $\mathrm{URad}(V + \{u\}) = \mathrm{URad}(V)$.
4. $\mathrm{URad}(\{\pm 1\}^n) = \mathrm{E}\epsilon^\top \epsilon = n$.
5. $\mathrm{URad}(\{(-1, \ldots, -1)^\top, (+1, \ldots, +1)^\top\}) = \mathrm{E}\left|\sum_{i=1}^{n} \epsilon_i\right| = O(\sqrt{n})$.

Main Result of This Section

The following theorem shows indeed we can use Rademacher complexity to replace the $\log|\mathcal{F}|$ term from finite-class bound with something more general.

Theorem (18.1)

Let $\mathcal{F}$ be given with $f(z) \in [a,b]$ a.s. $\forall f \in \mathcal{F}$.

1. *With probability $\geq 1-\delta$,*

$$\sup_{f\in\mathcal{F}} \mathrm{E}f(Z) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \leq \mathrm{E}_{\{Z_i\}_{i=1}^n}\left(\sup_{f\in\mathcal{F}}\left(\mathrm{E}f(Z) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right)\right) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

2. *With probability $\geq 1-\delta$,*

$$\mathrm{E}_{\{Z_i\}_{i=1}^n} \mathrm{URad}(\mathcal{F}_{|S}) \leq \mathrm{URad}(\mathcal{F}_{|S}) + (b-a)\sqrt{\frac{n\log(1/\delta)}{2}}.$$

3. *With probability $\geq 1-\delta$,*

$$\sup_{f\in\mathcal{F}}\left(\mathrm{E}f(Z) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right) \leq \frac{2}{n}\mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The Main Result of This Section

Remark

We can get the absolute value version of $\sup_{f\in\mathcal{F}}\left(\mathrm{E}f(Z)-\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\right)$ *by just replace* $\mathcal{F}$ *with* $-\mathcal{F}:=\{-f:f\in\mathcal{F}\}$*.*

Proof

The proof of this bound has many interesting points, and we will prove it later. It has these basic steps:

1. The *expected* uniform deviations can be upper bounded by the *expected* Rademacher complexity:
   a. The expected deviations are upper bounded by expected deviations between two finite samples. This is interesting since we could have reasonably defined generalization in terms of this latter quantity.
   b. These two-sample deviations are upper bounded by expected Rademacher complexity by introducing random signs.
2. We replace this difference in expectations with high probability bounds via a more powerful concentration inequality: McDiarmid's inequality.

Denote

$$Pf := \mathrm{E}_Z f(Z), \qquad \mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(Z_i), \qquad P_n g = \mathrm{E}_{\{Z_i\}_{i=1}^n} g(Z_1, \ldots, Z_n)$$

We will prove the Main Result by steps.

Symmetrization with a ghost sample:

Consider "ghost sample" $\{Z_i'\}_{i=1}^n$ to be iid draw from Z, and define P_n' and $\mathbb{P}_n'$ analogously.

Lemma (18.1)

$$P_n\left(\sup_{f\in\mathcal{F}}(P - \mathbb{P}_n)f\right) \leq P_n P_n'\left(\sup_{f\in\mathcal{F}}(\mathbb{P}_n' - \mathbb{P}_n)f\right).$$

Proof of Lemma 18.1

Fix any $\epsilon > 0$ and apx $\max f_\epsilon \in \mathcal{F}$, then

$$\begin{aligned} P_n\left(\sup_{f\in\mathcal{F}}(P-\mathbb{P}_n)f\right) &\leq P_n\big((P-\mathbb{P}_n)f_\epsilon\big)+\epsilon \\ \text{(IID sample)} &= P_n\left(P_n'\mathbb{P}_n' f_\epsilon f_\epsilon - \mathbb{P}_n f_\epsilon\right)+\epsilon \\ &= P_n' P_n(\mathbb{P}_n' f_\epsilon - \mathbb{P}_n f_\epsilon)+\epsilon \\ &\leq P_n' P_n\left(\sup_{f\in\mathcal{F}}(\mathbb{P}_n' - \mathbb{P}_n)f\right)+\epsilon. \end{aligned}$$

Results follows since $\epsilon > 0$ is arbitrary.

Remark

From Lemma 18.1, we know that we can instead work with two sample, i.e. the symmetrization with a ghost sample.

Symmetrization with Random Signs

Connecting two sample with Rademacher complexity by using random signs.

Lemma (18.2)

$$P_n P_n' \left(\sup_{f \in \mathcal{F}} (\mathbb{P}_n' - \mathbb{P}_n) f \right) \leq \frac{2}{n} P_n \, \mathrm{URad}(\mathcal{F}_{|S}).$$

Fix a vector $\epsilon \in \{\pm 1\}^n$, define $(U_i, U_i') = (Z_i, Z_i')$ if $\epsilon_i = 1$ and $(U_i, U_i') = (Z_i', Z_i)$ if $\epsilon_i = -1$. Then

$$
\begin{aligned}
\mathrm{E}_\epsilon P_n P_n' \left(\sup_{f \in \mathcal{F}} (\mathbb{P}_n' - \mathbb{P}_n) f \right) &= \mathrm{E}_\epsilon P_n P_n' \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \big(f(Z_i') - f(Z_i) \big) \right) \\
&= \mathrm{E}_\epsilon P_n P_n' \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(U_i') - f(U_i) \big) \right) \\
\big((Z_1, \ldots, Z_n') \stackrel{d}{=\!=} (U_1, \ldots, U_n') \big) &= \mathrm{E}_\epsilon P_n P_n' \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(Z_i') - f(Z_i) \big) \right) \\
&\le \mathrm{E}_\epsilon P_n P_n' \left(\sup_{f, f' \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(Z_i') - f'(Z_i) \big) \right) \\
&= \mathrm{E}_\epsilon P_n' \left(\sup_{f'} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(f(Z_i') \big) \right) \\
&\qquad + \mathrm{E}_\epsilon P_n \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \big(- f'(Z_i) \big) \right) \\
&= 2 P_n \frac{1}{n} \mathrm{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(Z_i) = 2 P_n \left(\frac{1}{n} \mathrm{URad}(\mathcal{F}_{|S}) \right).
\end{aligned}
$$

Generalization *with* concentration

Now, we will control the expected uniform deviations: $P_n \sup_{f \in \mathcal{F}} (P - \mathbb{P}_n) f$ with high probability bounds follows via the *McDiarmid Inequality*:

Theorem (McDiarmid Inequality)

Suppose $F : \mathbb{R}^n \to \mathbb{R}$ *satisfies "bound differences":* $\forall i \in \{1, \ldots, n\}$, $\exists c_i > 0$,

$$\sup_{z_1, \ldots, z_n, z_i'} \left| F(z_1, \ldots, z_i, \ldots, z_n) - F(z_1, \ldots, z_i', \ldots, z_n) \right| \leq c_i,$$

then with probability $\geq 1 - \delta$,

$$P_n F(Z_1, \ldots, Z_n) \leq F(Z_1, \ldots, Z_n) + \sqrt{\frac{\sum_{i=1}^n c_i^2}{2} \log(1/\delta)}.$$

The Proof of the Main Result

For the first result, denote $\mathbb{P}_n^* f := \frac{1}{n}\sum_{i=1}^n (f(Z_1) + \ldots + f(Z_i') + \ldots + f(Z_n))$, the result comes from the McDiarmid Inequality and

$$\left| \sup_{f\in\mathcal{F}} (P - \mathbb{P}_n) f - \sup_{f\in\mathcal{F}} (P - \mathbb{P}_n^*) f \right| \leq \sup_{f\in\mathcal{F}} \frac{|f(Z_i') - f(Z_i)|}{n} \leq \frac{b-a}{n}.$$

For the second result, denote $S' = \{Z_1, \ldots, Z_i', \ldots, Z_n\}$, the result similarly comes from

$$\left| \mathrm{URad}(\mathcal{F}_{|S}) - \mathrm{URad}(\mathcal{F}_{|S'}) \right| = \left| \mathrm{E}_\epsilon \big(\epsilon_i f(Z_i) - \epsilon_i f(Z_i') \big) \right| \leq (b-a).$$

Then the third result comes from the Lemma 18.1, Lemma 18.2, and the second result in Theorem.

Example: Logistic Regression

For Logistic Regression:

$$\ell(yf(z)) := \log\big(1 + \exp(-yf(x))\big), \qquad |\ell'| \le 1, \qquad \mathcal{F} := \{w \in \mathbb{R}^d : \|w\| \le B\}$$

$$(\ell \circ \mathcal{F})_{|S} = \Big\{ \big(\ell(Y_1 w^\top X_1), \ldots, \ell(Y_n w^\top X_n)\big) : \|w\| \le B \Big\},$$

$$\mathcal{R}_\ell(w) := \mathrm{E}\ell(Y w^\top X), \qquad \widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i w^\top X_i).$$

Our Goal: Control $\mathcal{R}_\ell - \widehat{\mathcal{R}}_\ell$ over $\mathcal{F}$ through $\mathrm{URad}\big((\ell \circ \mathcal{F})_{|S}\big)$.

Step 1: "Peeling" off ℓ.

Step 2: Rademacher complexity of linear predictors.

Example: Logistic Regression; Step 1

Lemma (18.3, Contraction Property)

Let $\ell : \mathbb{R}^n \to \mathbb{R}$ *be a vector of univariate* L*-lipschitz functions. Then* $\mathrm{URad}(\ell \circ V) \leq L \cdot \mathrm{URad}(V)$.

From this lemma, we can get the following corollary from Theorem 18.1 3. immediately.

Corollary

Suppose ℓ *is* L*-lipschitz and* $\ell \circ \mathcal{F} \in [a, b]$ *a.s. Then with probability* $\geq 1 - \delta$, *every* $f \in \mathcal{F}$ *satisfies*

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + \frac{2L}{n} \mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The idea of the proof of Lemma 18.3 is to "de-symmetrize" and get a difference of coordinates to which we can apply the definition of L.

The Proof of Lemma 18.3

To start with,

$$
\begin{aligned}
&\mathrm{URad}(\ell \circ V) = \mathrm{E}_\epsilon \sup_{u \in V} \sum_{i=1}^{n} \epsilon_i \ell_i(u_i) \\
=&\mathrm{E}_{\epsilon_{2:n}} \mathrm{E}_{\epsilon_1} \left[\sup_{u_1 \in V_1} \epsilon_1 \ell_1(u_1) + \sup_{u_{2:n} \in V_{-1}} \sum_{i=2}^{n} \epsilon_i \ell_i(u_i) \right] \\
=&\mathrm{E}_{\epsilon_{2:n}} \left[\sup_{u_1 \in V_1} \frac{1}{2} \ell_1(u_1) + \sup_{u_1 \in V_1} \frac{-1}{2} \ell_1(u_1) + \frac{1}{2} \sup_{u_{2:n} \in V_{-1}} \sum_{i=2}^{n} \epsilon_i \ell_i(u_i) + \frac{1}{2} \sup_{w_{2:n} \in V_{-1}} \sum_{i=2}^{n} \epsilon_i \ell_i(w_i) \right] \\
=&\frac{1}{2} \mathrm{E}_{\epsilon_{2:n}} \sup_{u,w \in V} \left[\ell_1(u_1) - \ell_1(w_1) + \sum_{i=2}^{n} \epsilon_i \big(\ell_i(u_i) + \ell_i(w_i) \big) \right] \\
\leq&\frac{1}{2} \mathrm{E}_{\epsilon_{2:n}} \sup_{u,w \in V} \left[L|u_1 - w_1| + \sum_{i=2}^{n} \epsilon_i \big(\ell_i(u_i) + \ell_i(w_i) \big) \right] \\
=&\frac{1}{2} \mathrm{E}_{\epsilon_{2:n}} \sup_{u,w \in V} \left[L(u_1 - w_1) + \sum_{i=2}^{n} \epsilon_i \big(\ell_i(u_i) + \ell_i(w_i) \big) \right] = \mathrm{E}_\epsilon \sup_{u \in V} \left[L\epsilon_1 u_1 + \sum_{i=2}^{n} \epsilon_i \ell_i(u_i) \right].
\end{aligned}
$$

Then repeating this procedure for the other coordinates gives the bound in the lemma.

Example: Logistic Regression; Step 2

Theorem (18.3)

Collect sample $S := \{x_1, \ldots, x_n\}$ into rows of $X \in \mathbb{R}^{n\times d}$,

$$\mathrm{URad}\left(\{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\}_{|S}\right) \le B\|X\|_F.$$

Fix any $\epsilon \in \{\pm 1\}^n$, then

$$\sup_{\|w\|\le B} \sum_{i=1}^n \epsilon_i \langle w, x_i\rangle = \sup_{\|w\|\le B} \left\langle w, \sum_{i=1}^n \epsilon_i x_i \right\rangle = B\left\|\sum_{i=1}^n \epsilon_i x_i\right\|.$$

Since

$$\mathrm{E}\left\|\sum_{i=1}^n \epsilon_i x_i\right\| \le \mathrm{E}\sqrt{\left\|\sum_{i=1}^n \epsilon_i x_i\right\|^2} = \|X\|_F^2$$

we conclude this theorem.

Example: Logistic Regression

Suppose $\|w\|_2 \leq B$ and $\|x_i\|_2 \leq 1$. For logistic loss $\ell(z) := \log(1 + \exp(z))$, we have

$$\ell(\langle w, yx \rangle) \geq 0, \qquad \ell(\langle w, yx \rangle) \leq \begin{cases} \log 2 & \langle w, yx \rangle < 0 \\ \log 2 + \langle w, yx \rangle & \langle w, yx \rangle \geq 0 \end{cases} \leq \log 2 + B.$$

Combined with the previous results, we have with probability at least $1 - \delta$, every $w \in \mathbb{R}^d$ with $\|w\|_2 \leq B$ satisfies

$$\begin{aligned} \mathcal{R}_\ell(w) &\leq \widehat{\mathcal{R}}_\ell(w) + \frac{2}{n} \operatorname{URad}((\ell \circ \mathcal{F})_{|S}) + 3\big(\log 2 + B\big)\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq \widehat{R}_\ell(w) + \frac{2B\|X\|_F}{n} + 3\big(\log 2 + B\big)\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq \widehat{\mathcal{R}}_\ell(w) + \frac{2B + 3(B + \log 2)\sqrt{\log(2/\delta)/2}}{\sqrt{n}}. \end{aligned}$$

Some Basic Properties of Rademacher Complexity

We have given some properties of Rademacher Complexity in previous remarks when defining Rademacher Complexity. The following is other important properties.

Lemma (19.1)

1. $\mathrm{URad}(V) \geq 0$.
2. $\mathrm{URad}(cV + \{u\}) \leq |c| \, \mathrm{URad}(V)$.
3. $\mathrm{URad}\big(\mathrm{conv}(V)\big) = \mathrm{URad}(V)$.
4. *Let* $\{V_i\}_{i\geq 1}$ *be given with* $\sup_{u \in V_i} \langle u, \epsilon \rangle \geq 0$, $\forall \epsilon \in \{\pm 1\}^n$ *(e.g.* $V_i = -V_i$ *or* $0 \in V_i$*), then* $\mathrm{URad}\big(\cup_{i\geq 1} V_i\big) \leq \sum_{i\geq 1} \mathrm{URad}(V_i)$.
5. $\mathrm{URad}(V) = \mathrm{URad}(-V)$.

Proofs

1. Fix any $v \in V$, then $\mathrm{URad}(V) = \mathrm{E}_\epsilon \sup_{u \in V} \langle \epsilon, u \rangle \geq \mathrm{E}_\epsilon \langle \epsilon, v \rangle = 0$.

2. Let $\ell_i(r) = c \cdot r + u_i$, and it directly comes from Lemma 18.3, the contraction property of Rademacher complexity.

3. This follows since optimization over a polytope is achieved at a corner.

$$\begin{aligned}
\mathrm{URad}\left(\mathrm{conv}(V)\right) &= \mathrm{E}_\epsilon \sup_{k \geq 1,\, \alpha \in \Delta_k} \sup_{u_1, \ldots, u_k \in V} \left\langle \epsilon, \sum_{j=1}^{k} \alpha_j u_j \right\rangle \\
&= \mathrm{E}_\epsilon \sup_{k \geq 1,\, \alpha \in \Delta_k} \sum_{j=1}^{k} \alpha_j \sup_{u_j \in V} \langle \epsilon, u_j \rangle \\
&= \mathrm{E}_\epsilon \left(\sup_{k \geq 1,\, \alpha \in \Delta_k} \sum_{j=1}^{k} \alpha_j \right) \sup_{u \in V} \langle \epsilon, u \rangle = \mathrm{E}_\epsilon \sup_{u \in V} \langle \epsilon, u \rangle = \mathrm{URad}(V),
\end{aligned}$$

where $\Delta_k := \{x \in \mathbb{R}^k : x_i \geq 0,\ \|x\|_1 = 1\}$.

4. Using the condition

$$\mathrm{URad}\left(\cup_i V_i\right) = \mathrm{E}_\epsilon \sup_{u \in \cup_i V_i} \langle \epsilon, u \rangle = \mathrm{E}_\epsilon \sup_i \sup_{u \in V_i} \langle \epsilon, u \rangle \leq \mathrm{E}_\epsilon \sum_i \sup_{u \in V_i} \langle \epsilon, u \rangle = \sum_i \mathrm{URad}(V_i).$$

Two Rademacher Complexity Proofs for Deep NN

For a matrix $A_{m\times n} = (a_1, \ldots, a_n) \in \mathbb{R}^{m\times n}$, denote

$$\|A\|_{b,c} := \left\|(\|a_1\|_b, \ldots, \|a_n\|_b)^\top\right\|_c.$$

There are two bounds obtained by inductively peeling off layers.

1. One will depend on $\|W_i^\top\|_{1,\infty}$ (see [1]).
2. The other will depend on $\|W_i^\top\|_F$ (see [2]).

First Layer Peeling Proof: $(1, \infty)$ norm

Theorem (19.1, First Layer Peeling Proof: $(1, \infty)$ norm)

Let ρ*-Lipschitz activations* σ_i *satisfy* $\sigma_i(0) = 0$*, and*

$$\mathcal{F} = \left\{x \mapsto \sigma_L(W_L \sigma_{L-1}(\cdots \sigma_1(W_1 x) \cdots)) : \|W_i^\top\|_{1,\infty} \le B\right\}.$$

Then $\mathrm{URad}(\mathcal{F}_{|S}) \le \|X\|_{2,\infty} (2\rho B)^L \sqrt{2 \log d}$*, where the data matrix is*

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (x_{(1)}, \cdots, x_{(d)}) \in \mathbb{R}^{n \times d}.$$

Remark

1. $(\rho B)^L$ is roughly a Lipschitz constant of the networki according to ∞-norm bounded inputs, which is related to the "worst case" whereas ideally we want "average case".
2. The factor 2^L is not good and we will use Frobenius norm $\|W\|_F$ to remove it.

Proof

We will use induction to show the theorem. Let $\mathcal{F}_i$ denote functions computed by nodes in layer i.

Base case ($i = 0$):

$$\mathcal{F}_0 = \{x \mapsto x_j : j \in [d]\},$$

since Y is one-dimension, and hence

$$(\mathcal{F}_0)_{|S} = \left\{ \left((x_1)_j, \ldots, (x_n)_j\right)^\top : j \in [d] \right\}.$$

Therefore, by the Massart Finite Lemma ($\mathrm{URad}(V) \le \sup_{u \in V} \|u\|_2 \sqrt{2 \log |V|}$),

$$\begin{aligned}\mathrm{URad}\left((\mathcal{F}_0)_{|S}\right) &\le \left(\max_{j \in [d]} \|((x_1)_j, \ldots, (x_n)_j)^\top\|_2 \right) \sqrt{2 \log d} \\ &= \|X\|_{2,\infty} \sqrt{2 \log d} = \|X\|_{2,\infty} (2\rho B)^0 \sqrt{2 \log d}.\end{aligned}$$

Proof

Inductive Step: Applying both Lipschitz peeling and the preceding multi-part lemma,

$$\operatorname{URad}\left((\mathcal{F}_{i+1})_{|S}\right) = \operatorname{URad}\left(\left\{x \mapsto \sigma_{i+1}\left(W_{i+1}^{\top} g(x)\right) : g \in (\mathcal{F}_i)\right\}_{|S}\right)$$

$$\overset{\text{something like Lemma 17 [3]?}}{=} \operatorname{URad}\left(\left\{x \mapsto \sigma_{i+1}\left(\|W_{i+1}^{\top}\|_{1,\infty} g(x)\right) : g \in \operatorname{conv}(-\mathcal{F}_i \cup \mathcal{F}_i)\right\}_{|S}\right)$$

$$\overset{\text{Lemma 18.3}}{\leq} (\rho B) \cdot \operatorname{URad}\left(\left(\operatorname{conv}(-\mathcal{F}_i \cup \mathcal{F}_i)\right)_{|S}\right)$$

$$\overset{\text{Lemma 19.1 3.}}{=} (\rho B) \cdot \operatorname{URad}\left((-\mathcal{F}_i \cup \mathcal{F}_i)_{|S}\right)$$

$$\overset{\text{Lemma 19.1 4.}}{\leq} 2\rho B \cdot \operatorname{URad}\left((\mathcal{F}_i)_{|S}\right) \leq \cdots \leq (2\rho B)^{i+1} \|X\|_{2,\infty} \sqrt{2 \log d},$$

where the last step is by the fact $0 = \sigma(\langle 0, F(x)\rangle) \in \mathcal{F}_i$.

Two Rademacher Complexity Proofs for Deep NN

Theorem (19.2)

Let 1-Lipschitz positive homogeneous activation σ_i be given, and

$$\mathcal{F} := \left\{ x \mapsto \sigma_L(W_L \sigma_{L-1}(\cdots \sigma_1(W_1 x) \cdots)) : \|W_i\|_F \leq B \right\}.$$

Then

$$\mathrm{URad}(\mathcal{F}_{|S}) \leq B^L \|X\|_F \left(1 + \sqrt{2L \log 2}\right).$$

Remark

We do not include 2^L! the main proof trick is to replace E_ϵ with $\log \mathrm{E}_\epsilon \exp$, and 2^L now appears inside $\log$.

Proof Steps

1. "Lipschitz Peeling" with $\exp$ inside E_ϵ;
2. Use Massart Finite Lemma to deal with the base case of layer peeling;

An Important Lemma

We will use an important lemma to Lipschitz peel.

Lemma (Extra 2, see Eq. 4.20 in [4])

Let $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ *is a contraction (univariate* 1*-Lipschitz) with* $\varphi(0) = 0$*. If* $G : \mathbb{R} \to \mathbb{R}$ *is convex and increasing, then*

$$\mathrm{E}G\left(\sup_{u \in V} \sum_{i=1}^{n} \epsilon_i \varphi_i(u_i)\right) \leq \mathrm{E}G\left(\sup_{u \in V} \sum_{i=1}^{n} \epsilon_i u_i\right).$$

Lipschitz peeling bound.

We will use the following refined Lipschitz peeling bound.

Lemma (19.2)

Let $\ell : \mathbb{R}^n \to \mathbb{R}^n$ be a vector of univariate ρ-Lipschitz functions with $\ell_i(0) = 0$. Then

$$\mathrm{E}_\epsilon \exp\left(\sup_{u\in V}\sum_i \epsilon_i \ell_i(u_i)\right) \leq \mathrm{E}_\epsilon \exp\left(\rho \sup_{u\in V}\sum_i \epsilon_i u_i\right).$$

Proof

Note that $\ell_i(\rho^{-1}\cdot)$ is a contraction and $G(x) := \exp(x)$ is convex and increasing,

$$\begin{aligned}\mathrm{E}_\epsilon \exp\left(\sup_{u\in V}\sum_i \epsilon_i \ell_i(u_i)\right) &= \mathrm{E}_\epsilon \exp\left(\sup_{\rho u\in \rho V}\sum_i \epsilon_i \ell_i(\rho^{-1}\cdot \rho u_i)\right) \\ &\leq \mathrm{E}_\epsilon \exp\left(\sup_{\rho u\in \rho V}\sum_i \epsilon_i \cdot \rho u_i\right) = \mathrm{E}_\epsilon \exp\left(\rho \sup_{u\in V}\sum_i \epsilon_i u_i\right).\end{aligned}$$

Frobenius Norm for sub-Gaussian

And the peeling proof will end with a term $\mathrm{E}\exp(t\|X^\top\epsilon\|)$, and we'll optimize the t o get the final bound: we actually are proving $\|X^\top\epsilon\|$ is sub-Gaussian.

Lemma (19.3)

$\mathrm{E}\|X^\top\epsilon\|_2 \le \|X\|_F$, *and* $\|X^\top\epsilon\|_2 - \mathrm{E}\|X^\top\epsilon\|_2 \sim \mathrm{subG}(\|X\|_F^2)$.

First,

$$\mathrm{E}\|X^\top\epsilon\|_2 \le \sqrt{\mathrm{E}\|X^\top\epsilon\|_2^2} = \sqrt{\sum_{j=1}^{d} \mathrm{E}(x_{(j)}^\top\epsilon)^2} = \sqrt{\sum_{j=1}^{d} \|x_{(j)}\|^2} = \|X\|_F.$$

This is the first result in the Lemma.

Proof of Lemma 19.3: Method 1

Suppose ϵ and ϵ' only differ on ϵ_i,

$$
\begin{aligned}
\sup_{\epsilon,\,\epsilon'} \left| \|X^\top \epsilon\|_2 - \|X^\top \epsilon'\|_2 \right|^2 &\le \sup_{\epsilon,\,\epsilon'} \|X^\top(\epsilon - \epsilon')\|_2^2 = \sup_{\epsilon,\,\epsilon'} \sum_{j=1}^{d} \left(x_{(j)}^\top (\epsilon - \epsilon') \right)^2 \\
&= \sup_{\epsilon,\,\epsilon'} \sum_{j=1}^{d} \left(x_{ij}(\epsilon_i - \epsilon_i') \right)^2 \le 4\|x_i\|_2^2.
\end{aligned}
$$

And hence,

$$
\sup_{\epsilon,\,\epsilon'} \left| \|X^\top \epsilon\|_2 - \|X^\top \epsilon'\|_2 \right| \le 2\|x_i\|_2.
$$

By McDiarmid Inequality, we have

$$
\mathrm{P}\left(\left| \|X^\top \epsilon\|_2 - \mathrm{E}\|X^\top \epsilon\|_2 \right| \ge \epsilon \right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} (2\|x_i\|_2)^2} \right) \le 2\exp\left(-\frac{\epsilon^2}{2\|X\|_F^2} \right),
$$

which implies $\|X^\top \epsilon\|_2 - \mathrm{E}\|X^\top \epsilon\|_2$ is sub-Gaussian with variance proxy $c\|X\|_F^2$ with some positive $c > 0$ is free of X.

Proof of Lemma 19.3: Method 2

Recall the Hoeffding's Lemma,

Lemma (Extra 3, Hoeffding's Lemma)

Suppose $\xi \in [a, b]$ *a.s., then for all* $\lambda \in \mathbb{R}$*,*

$$\mathrm{E}\exp\left(\lambda(\xi - \mathrm{E}\xi)\right) \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

Specially, we have $(\xi - \mathrm{E}\xi) \sim \mathrm{subG}\left((b-a)^2/4\right)$.

Define

$$Y_i := \mathrm{E}\left[\|X^\top \epsilon\|_2 \,|\, \epsilon_1, \ldots, \epsilon_i\right], \qquad D_i := Y_i - Y_{i-1},$$

whereby $Y_n - Y_0 = \sum_{i=0}^n D_i$.

Proof of Lemma 19.3: Method 2

Since $\sup_{\epsilon,\,\epsilon'} \big|\|X^\top\epsilon\|_2 - \|X^\top\epsilon'\|_2\big| \le 2\|x_i\|_2$, $D_i \in \big[-2\|x_i\|_2, 2\|x_i\|_2\big]$, and then $D_i - \mathrm{E}D_i \in \mathrm{subG}(\|x_i\|_2^2)$. And by the sub-Gaussian Property Lemma (see p8), we have

$$\|X^\top\epsilon\|_2 - \mathrm{E}\|X^\top\epsilon\|_2 = \sum_{i=1}^n (D_i - \mathrm{E}D_i) \sim \mathrm{subG}\left(\sum_{i=1}^n \|x_i\|_2^2\right) = \mathrm{subG}(\|X\|_F^2).$$

Details

$$\begin{aligned}
D_i &= \mathrm{E}\Big[\|X^\top\epsilon\|_2 \,\Big|\, \epsilon_1,\ldots,\epsilon_i\Big] - \mathrm{E}\Big[\|X^\top\epsilon\|_2 \,\Big|\, \epsilon_1,\ldots,\epsilon_{i-1}\Big] \\
&= \mathrm{E}\Big[\|X^\top\epsilon\|_2 \,\Big|\, \epsilon_1,\ldots,\epsilon_i\Big] - \mathrm{E}_{\epsilon_i'}\Big[\mathrm{E}\big[\|X^\top\epsilon'\|_2 \,\big|\, \epsilon_1,\ldots,\epsilon_i'\big]\Big] \\
&= \mathrm{E}_{\epsilon_i'}\Big[\mathrm{E}\big[\|X^\top\epsilon\|_2 \,\big|\, \epsilon_1,\ldots,\epsilon_i\big]\Big] - \mathrm{E}_{\epsilon_i'}\Big[\mathrm{E}\big[\|X^\top\epsilon'\|_2 \,\big|\, \epsilon_1,\ldots,\epsilon_i'\big]\Big] \\
&= \mathrm{E}_{\epsilon_i'}\Big[\mathrm{E}\big[\|X^\top\epsilon\|_2 - \|X^\top\epsilon'\|_2 \,\big|\, \epsilon_1,\ldots,\epsilon_i,\epsilon_i'\big]\Big].
\end{aligned}$$

Proof of Theorem 19.2

Let X_i denote the output of layer i, i.e.

$$X_0 := X \qquad \text{and} \qquad X_i := \sigma_i(X_{i-1}W_i^\top).$$

Step 1: "Lipschitz Peeling".

Define $\sigma := \sigma_i$, $Y := X_{i-1}$, $V := W_i$, $\widetilde{V}$ has ℓ_2-normalized rows, and w be all parameters acrosss all layers. Introduce u denote an arbitrary unit norm vector, by the property of $\sigma(\cdot)$ and $\|W_i\|_F \leq B$,

$$\begin{aligned}
\sup_w \|\epsilon^\top X_i\|_2^2 &= \sup_w \sum_j \left(\epsilon^\top \sigma(YV^\top)_{:j}\right)^2 = \sup_w \sum_j \left(\epsilon^\top \sigma(YV_{j:}^\top)\right)^2 \\
&= \sup_w \sum_j \left(\epsilon^\top \sigma(\|V_{j:}\| Y \widetilde{V}_{j:}^\top)\right)^2 = \sup_w \sum_j \|V_{j:}\|_2^2 \left(\epsilon^\top \sigma(Y\widetilde{V}_{j:}^\top)\right)^2 \\
&\leq \sup_w \sum_j \|V_{j:}\|_2^2 \sup_u \left(\epsilon^\top \sigma(Yu)\right)^2 = \sup_{w,u} \|V\|_F^2 \left(\epsilon^\top \sigma(Yu)\right)^2 \\
&\leq \sup_{w,u} B^2 \left(\epsilon^\top \sigma(Yu)\right)^2.
\end{aligned}$$

Proof of Theorem 19.2

Then

$$
\begin{aligned}
\mathrm{E}_\epsilon \exp\left(t\sqrt{\sup_w \|\epsilon^\top X_i\|_2^2}\right) &\le \mathrm{E}_\epsilon \exp\left(t\sqrt{\sup_{w,u} B^2(\epsilon^\top\sigma(Yu))^2}\right) \\
&\le \mathrm{E}_\epsilon \sup_{w,u} \exp\left(tB\epsilon^\top\sigma(Yu)\right) + \mathrm{E}_\epsilon \sup_{w,u} \exp\left(-tB\epsilon^\top\sigma(Yu)\right) \\
&= \mathrm{E}_\epsilon 2 \sup_{w,u} \exp\left(tB\epsilon^\top\sigma(Yu)\right) \\
&\le \mathrm{E}_\epsilon 2 \sup_{w,u} \exp\left(tB\epsilon^\top Yu\right) \\
&\le \mathrm{E}_\epsilon 2 \sup_{w} \exp\left(tB\|\epsilon^\top Y\|_2\right) \\
&\le \cdots \le \mathrm{E}_\epsilon 2^i \sup_{w} \exp\left(tB^i\|\epsilon^\top X_0\|_2\right).
\end{aligned}
$$

Why "$\le$"

$$
(\epsilon^\top Yu)^2 = \mathrm{tr}\left(\epsilon^\top Yuu^\top Y^\top\epsilon\right) = \mathrm{tr}\left(Y^\top\epsilon\epsilon^\top Yuu^\top\right) \le \mathrm{tr}\left(Y^\top\epsilon\epsilon^\top Y\right)\mathrm{tr}\left(uu^\top\right) = \|\epsilon^\top Y\|_2^2.
$$

Proof of Theorem 19.2

Step 2: Setting $\mu := \mathrm{E}\|X_0^\top \epsilon\|_2$,

$$\begin{aligned}
\mathrm{URad}(\mathcal{F}_{|S}) &= \mathrm{E}\sup_w \epsilon^\top X_L = \mathrm{E}\frac{1}{t}\log\sup_w \exp(t\epsilon^\top X_L) \\
&\overset{\text{Jensen Inequality}}{\leq} \frac{1}{t}\log \mathrm{E}\sup_w \exp\left(t|\epsilon^\top X_L|\right) \leq \frac{1}{t}\log \mathrm{E}\exp\left(t\sqrt{\sup_w \|\epsilon^\top X_L\|_2^2}\right) \\
&\overset{\text{Step 1}}{\leq} \frac{1}{t}\log \mathrm{E}2^L\exp\left(tB^L\|\epsilon^\top X_0\|_2\right) = \frac{1}{t}\log \mathrm{E}2^L\exp\left(tB^L(\|\epsilon^\top X_0\|_2 - \mu + \mu)\right) \\
&\overset{\text{Lemma 19.3}}{\leq} \frac{1}{t}\log\left[2^L\exp\left(t^2B^{2L}\|X\|_F^2/2 + tB^L\mu\right)\right] \\
&= \frac{L\log 2}{t} + \frac{tB^{2L}\|X\|_F^2}{2} + B^L\|X\|_F,
\end{aligned}$$

whereby the final bound follows with the minimizing choice

$$t := \sqrt{\frac{2L\log 2}{B^{2L}\|X\|_F^2}}.$$

Reference I

[1] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[2] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. "Size-independent sample complexity of neural networks". In: *Conference On Learning Theory*. PMLR. 2018, pp. 297–299.

[3] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "Norm-based capacity control in neural networks". In: *Conference on Learning Theory*. PMLR. 2015, pp. 1376–1401.

[4] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Thank You!